Comparative Evaluation of Chunking Strategies and Search Parameters in RAG Pipelines

Sachin Koirala

August 12, 2025

Abstract

This report presents a detailed comparison between **Semantic Chunking** and **Recursive Split Chunking** for Retrieval-Augmented Generation (RAG) pipelines, as well as between `SearchParams(hnsw_ef=128)` and `SearchParams(exact=True)` in vector similarity search using Qdrant. The analysis covers accuracy, precision, recall, F1-score, and latency, with theoretical expectations grounded in prior research and an evaluation plan inspired by Heliya Hasani's methodology for measuring chunking efficiency. This document serves both as a conceptual guide and as a template for reporting empirical results.

1 Introduction

Retrieval-Augmented Generation (RAG) pipelines rely heavily on effective text chunking and optimal vector search configurations. The chunking method determines how documents are split into retrievable units, while search parameters influence how efficiently relevant chunks are retrieved. The performance of these components directly affects downstream LLM answer quality.

This report compares:

1. **Semantic Chunking** vs **Recursive Split Chunking**
2. `hnsw_ef=128` vs `exact=True` search

based on expected performance characteristics, drawing from [1, 2] and related literature.

2 Background

2.1 Semantic Chunking

Semantic chunking groups text into chunks based on topical or meaning boundaries. Chunks tend to be coherent and contextually relevant, improving precision but sometimes reducing recall if relevant information spans multiple chunks.

2.2 Recursive Split Chunking

Recursive split chunking uses fixed-size chunks (tokens or characters) with optional overlaps. It is simple to implement and guarantees higher recall due to overlaps, but may reduce precision by introducing irrelevant content.

2.3 HNSW and Search Parameters

Qdrant's HNSW index supports:

- `hnsw_ef`: The number of candidate nodes explored during query time; higher values improve recall and accuracy at the cost of latency.
- `exact=True`: Performs an exhaustive similarity search, yielding ground-truth results but with significantly higher latency.

3 Methodology

3.1 Evaluation Metrics

We use:

- **Token-level IoU** (Intersection over Union)
- **Precision, Recall, F1-score** (token-level and chunk-level)
- **End-to-end Accuracy** (final answer correctness)
- **Latency** (retrieval and prompt processing time)
- **Indexing Cost** (time and compute to chunk + embed documents)

3.2 Experimental Setup

1. Prepare a labeled QA dataset with ground-truth answer spans.
2. Implement both chunking strategies.
3. Index with Qdrant using multiple search configurations (`hnsw_ef` values and `exact=True`).
4. Run retrieval + LLM answer generation for all configurations.
5. Collect metrics and analyze results statistically.

4 Comparative Analysis (Expected Results)

4.1 Chunking Methods

Table 1: Expected Performance of Chunking Methods

Metric	Semantic	Recursive
Accuracy	High	Medium
Precision	High	Low–Medium
Recall	Medium	High
F1-score	High	Medium
Latency	Medium	High
Indexing Cost	High	Low

4.2 Search Parameters

Table 2: Expected Performance of Search Parameters

Metric	hnsw_ef=128	exact=True
Accuracy/Recall	High	Highest
Precision	High	Highest
F1-score	High	Highest
Latency	Low–Med	Very High
Use Case	Production	Benchmarking

5 Proposed Experiment Plan

The evaluation plan follows Hasani's token-level metric approach [1]:

1. Test chunk sizes (e.g., 200–400 tokens) and overlaps (10–50%) for both strategies.
2. Test search parameters: `hnsw_ef` $\in \{16, 64, 128, 256\}$ and `exact=True`.
3. Measure IoU, Precision, Recall, F1, and latency per configuration.
4. Perform statistical significance tests to validate differences.

6 Conclusion and Recommendations

Based on theory and prior work:

- Use **Semantic Chunking + tuned** `hnsw_ef` for production RAG pipelines.
- Use `exact=True` to establish ground-truth retrieval performance for benchmarking.
- Apply small overlaps or hierarchical linking to semantic chunks to mitigate recall loss.

References

[1] H. Hasani, "How to measure chunking efficiency in RAG pipelines," *Medium*, 2025. Available: https://medium.com/@heliyahasani/how-to-measure-chunking-efficiency-in-rag-pipelines-0bb58499aa5a

[2] Qdrant Documentation, "Search parameters and HNSW index tuning," Available: https://qdrant.tech/documentation/